Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (500 word limit)

Key Decisions:

Answer these questions

1. What decisions needs to be made?

We need to decide if we are going to push through sending out the catalogues to the 250 new customers given the likelihood of purchasing anything, possible expected sales should they purchase, gross margin on those sales and the cost of printing. Per management, we will push through sending the catalogues if profits exceed $10,000.

2. What data is needed to inform those decisions?

We would first need the likelihood that the new customer will respond to the catalogue, which is given. We then need to predict sales given customer attributes. For this, we will use data on existing customers. Next, we would multiply Expected Sales with gross margin and deduct the cost of printing per catalogue to obtain the profit.

Step 2: Analysis, Modeling, and Validation

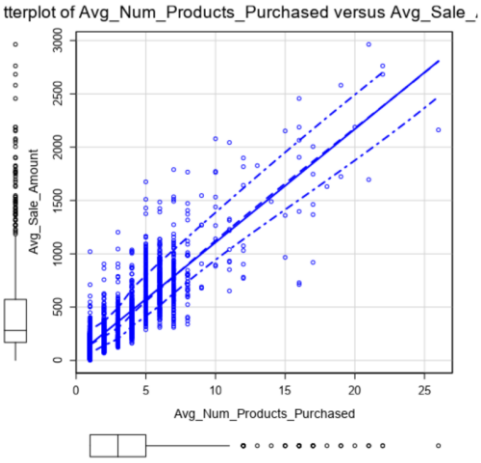
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

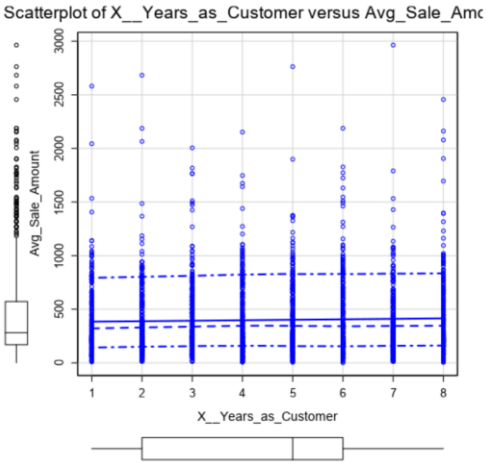
Important: Use the p1-customers.xlsx to train your linear model.

At the minimum, answer these questions:

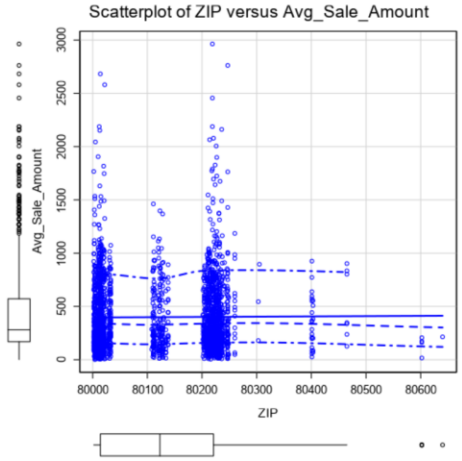
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

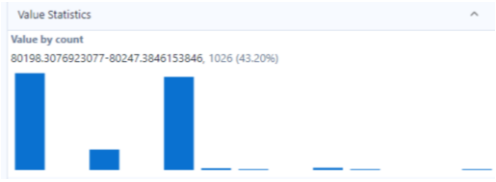
Average Number of Products purchased appears to be a strong predictor based on the linear relationship seen below in the scatterplot, whereas number of years as a customer is not.



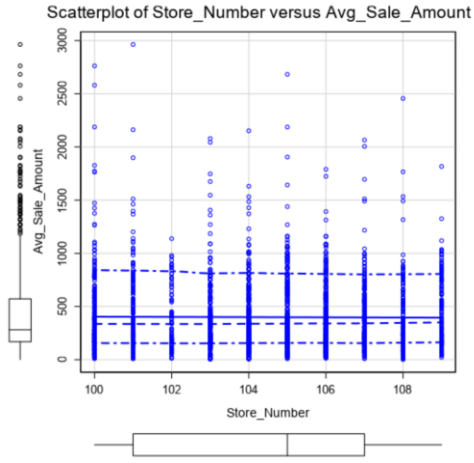


While the relationship is not linear since ZIP code is a categorical variable in numeric format, it appears that most sales are clumped in zip codes 80000-80200. However, from the distribution of values, most of the dataset have zip codes <80200.





No distinction can be observed for the Store Number.



City is a categorical variable so we can do univariate analysis with individual regression instead of a plot. Using 'city' as predictor variable, the model r-squared is very low which indicates that this is an unlikely predictor as well. There is a low p-value for Commerce City, however, this is not even a top city in terms of count as well.

Multiple R-squared: 0.008008,

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	386.087	21.67	17.81399	< 2.2e-16	***
CityAurora	18.755	26.55	0.70630	0.48007	
CityBoulder	154.103	197.85	0.77890	0.43612	
CityBrighton	-291.157	241.83	-1.20398	0.22872	
CityBroomfield	7.409	37.39	0.19816	0.84294	
CityCastle Pines	-193.877	241.83	-0.80171	0.4228	
CityCentennial	-13.816	44.24	-0.31230	0.75484	
CityCommerce City	296.728	109.87	2.70065	0.00697	**
CityDenver	18.551	24.99	0.74237	0.45794	
CityEdgewater	76.875	100.69	0.76349	0.44525	
CityEnglewood	-9.806	50.41	-0.19450	0.8458	
CityGolden	-12.719	81.09	-0.15685	0.87538	
CityGreenwood Village	-60.038	93.58	-0.64157	0.52121	
CityHenderson	-171.697	341.31	-0.50305	0.61498	
CityHighlands Ranch	4.904	74.26	0.06604	0.94735	
CityLafayette	-41.955	153.86	-0.27267	0.78513	
CityLakewood	31.652	31.69	0.99872	0.31803	
CityLittleton	-9.727	45.62	-0.21322	0.83118	
CityLone Tree	468.783	341.31	1.37348	0.16973	
CityLouisville	-37.619	171.68	-0.21912	0.82658	
CityMorrison	126.608	130.55	0.96977	0.33226	

Top Values

Denver	750
Aurora	493
Arvada	247
Lakewood	217
Broomfield	125
Westminster	85
Centennial	78
Littleton	72
Englewood	56
Wheat Ridge	54

Customer segment is also a categorical variable so we perform the same exercise as above and it turns out this is a highly predictive variable.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	682.7	8.354	81.72	< 2.2e-16	***
Customer_SegmentLoyalty Club Only	-286.3	11.372	-25.18	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	391.5	15.732	24.89	< 2.2e-16	***
Customer_SegmentStore Mailing List	-525.3	10.045	-52.30	< 2.2e-16	***

As such, our likely predictors are average number of products purchased and customer segment.

2. Explain why you believe your linear model is a good model. You must justify your reasoning using the statistical results that your regression model created. For each variable you selected, please justify how each variable is a good fit for your model by using the p-values and R-squared values that your model produced.

The resulting model is a good model as it has an R-squared> 0.7 at 0.8369, and all the predictor variables have p-values <0.05, which indicates statistical significance.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	303.46	10.576	28.69	< 2.2e-16	***
Customer_SegmentLoyalty Club Only	-149.36	8.973	-16.65	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	281.84	11.910	23.66	< 2.2e-16	***
Customer_SegmentStore Mailing List	-245.42	9.768	-25.13	< 2.2e-16	***
Avg_Num_Products_Purchased	66.98	1.515	44.21	< 2.2e-16	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 137.48 on 2370 degrees of freedom
Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366
F-statistic: 3040 on 4 and 2370 degrees of freedom (DF), p-value < 2.2e-16

3. What is the best linear regression equation based on the available data? Each coefficient should have no more than 2 digits after the decimal (ex: 1.28)

Important: The regression equation should be in the form:

*Y = Intercept + b1 * Variable_1 + b2 * Variable_2 + b3 * Variable_3……*

For example: Y = 482.24 + 28.83 * Loan_Status – 159 * Income + 49 (If Type: Credit Card) – 90 (If Type: Mortgage) + 0 (If Type: Cash)

Note that we **must** include the 0 coefficient for the type Cash.

Note: For students using software other than Alteryx, if you decide to use Customer Segment as one of your predictor variables, please set the base case to Credit Card Only.

Y= 303.46 + 66.98 * Avg_Num_Products_Purchases – 149.36 (If Customer_Segment: Loyalty Club Only) + 281.84 (If Customer_Segment: Loyalty Club and Credit Card) – 245.42 (If Customer_Segment: Mailing List) + 0 (If Customer_Segment: Credit Card Only)

Step 3: Presentation/Visualization

Use your model results to provide a recommendation. (500 word limit)

At the minimum, answer these questions:

1. What is your recommendation? Should the company send the catalog to these 250 customers?

The expected profit of $21,987.44 is ~double that of management's hurdle. As such, it is recommended to send out the catalog to the 250 customers.

2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)

We first used data on existing customers to create a model for **average sales**. Based on statistical results, the identified predictors of average sales are customer segment and average number of orders historically. We then applied this model to the new customers to obtain the average sales. However, to factor in the probability that they may or may not purchase as they receive the catalog, we had to multiple average sales to the probability that they will act on the catalog to get **expected sales**. Since we operate on a 50% margin for the items in the catalog, we compute **gross profit** using Expected Sales x 50% and subtract the cost of printing the catalog of $6.50/ea to obtain net **profit**.

Since this net profit is above the hurdle of $10,000 indicated by management, it is worthwhile to send out the catalog to the new customers.

3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?

The expected profit of $21,987.44 is ~double that of management's hurdle.

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.